STAT 651 Chapter 2.1–2.3

Zepu Zhang
December 16, 2010

If X is a rv, then a function of it, say $Y = g(X)$, is also a rv. Suppose we already know about X, then what do we know about Y? Since the behavior of Y is determined by the function g, all properties of Y can be studied through X and g. Essentially, the characteristics of Y is known. BUT, they may not be tractable. Nice, analytical forms may not exist for certain characteristics of Y.

The section 2.1 studies probabilistic descriptions about Y. We start with the general characteristics including sample space and probability function, and move on to more analytical descriptions including cdf and pmf/pdf. We'll start to get a feel with the examples in this section that many common distributions are connected to each other—a transform of a rv from one common distribution may lead to a rv that is distributed according to another common type.

We'll proceed from the completely general to more special, more tractable cases: general rv, discrete, continuous, continuous and monotone, continuous and piecewise monotone.

1 General random variables

Let X be a rv with sample space $\mathcal{X}$ and a defined probability function. (Remember, the probability function specifies the probability of the event $X \in A$ for any set A in the sigma algebra defined on $\mathcal{X}$.)

Let $Y = g(X)$ be a function of X:

$$g(x):\ \mathcal{X} \to \mathcal{Y}$$

The **sample space** $\mathcal{Y}$ is determined by $\mathcal{X}$ and g:

$$\mathcal{Y} = \{y :\ y = g(x), x \in \mathcal{X}\}$$

Example $\mathcal{X} = \{1, 2, 3, \dots\}$. $y = g(x) = x^2$.

Example $\mathcal{X} = (-\infty, \infty)$. $y = g(x) = x^2$.

To find the probability function of Y, we need to be able to go backwards: for event $A \subset \mathcal{Y}$, we need to find the event

$B \subset \mathcal{X}$ that "corresponds" to A, and find $P(B)$ based on the probability function of X. To do this, we need the "inverse" of the function g.

Let g^{-1} be an "inverse mapping" associated with g that maps from subsets of $\mathcal{Y}$ to subsets of $\mathcal{X}$. That is,

$$g^{-1}(A) = \{x \in \mathcal{X} : \; g(x) \in A\}$$

where $A \subset \mathcal{Y}$.

With g^{-1}, we can write the **probability function** of Y:

$$P(Y \in A) = P\big(X \in g^{-1}(A)\big)$$

Note the sets A and $g^{-1}(A)$ need not have the same "number" of elements. There are two situations for the function $g(x)$:

- g is one-to-one
  then g^{-1} is also one-to-one.
- g is many-to-one
  then g^{-1} is one-to-many.

(As a <u>function</u>, g can't be one-to-many or many-to-many.)

2 Discrete random variables

If X is a discrete rv, then $\mathcal{X}$ is countable. Then $\mathcal{Y}$ is a countable set. Thus Y is also a discrete rv.

In this case, the pmf of Y is easy to find:

$$y_Y(y) = P(Y = y) = \sum_{x \in g^{-1}(y)} f_X(x)$$

where $g^{-1}(y)$ means the same thing as $g^{-1}\big(\{y\}\big)$. Finding $f_Y(Y)$ entails identifying the elements of the set $g^{-1}(y)$, and then summing up the probabilities.

Example 2.1.1 (page 48).

3 cdf

In general,

$$F_Y(y) = P_Y(Y \leq y) = P_X\Big(g^{-1}\big((-\infty, y]\big)\Big)$$

For an arbitrary function g, the set $g^{-1}\big((-\infty, y]\big)$ may well contain multiple disjoint intervals (or even more complicated). Then the probability on the right-hand side may not be easy to find (or at least there may not be a simple formula).

Things are simpler if the function g is monotone.

Theorem 2.1.3 (page 51). Let $\mathcal{X} = \{x : \ f_X(x) > 0\}$ and $\mathcal{Y} = \{y : f_Y(x) > 0\}$.

- If g is increasing on $\mathcal{X}$, then
$$F_Y(y) = P(Y \le y) = P\big(X \le g^{-1}(y)\big) = F_X\big(g^{-1}(y)\big)$$
for $y \in \mathcal{Y}$.
- If g is decreasing on $\mathcal{X}$, then
$$F_Y(y) = P(Y \le y) = P\big(X \ge g^{-1}(y)\big) = 1 - F_X\big(g^{-1}(y)\big)$$
for $y \in \mathcal{Y}$.

Note 1. Note the modified definition of $\mathcal{X}$ and $\mathcal{Y}$. Now they only contain the values at which the pdf is positive. They are called the support sets of the rv's.
2. The theorem also holds for discrete X. The reason that we don't emphasize this is because the discrete case is really simple, and one may not need to use such "theorems".

Example 2.1.4 (page 51).

4 Continuous random variables

This section introduces two very useful results: theorems 2.1.5 and 2.1.10.

Example 2.1.4 continued: find the pdf.

Example 2.1.7 (page 52).

Theorem 2.1.5 (page 51). Deriving pdf of $Y = g(X)$ from pdf of X.

Conditions:
(1) g is monotone. (Then theorem 2.1.3 can be used. More on this below.)
(2) f_X is continuous on $\mathcal{X}$. (Then we can use $f_X = F'_X$.)
(3) $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. (Then Y has a continuous pdf.)

Let's try to understand this theorem.

The monotonicity of g is key. Such a transformation is one-to-one and onto from $\mathcal{X}$ to $\mathcal{Y}$. The transformation g uniquely pairs x's and y's. (See middle paragraph on page 50.) Let's suppose g is increasing. Then the relation

$$a < Y < b$$

is equivalent to

$$g^{-1}(a) < X < g^{-1}(b).$$

Therefore the two (equivalent) events have identical probability:

$$P(a < Y < b) = P\big(g^{-1}(a) < X < g^{-1}(b)\big).$$

In particular, let's do this:

$$\begin{aligned}
f_Y(y) &= \lim_{\delta \to 0} \frac{F_Y(y+\delta) - F_Y(y)}{\delta} \\
&= \lim_{\delta \to 0} \frac{F_X\big(g^{-1}(y+\delta)\big) - F_X\big(g^{-1}(y)\big)}{\delta} \\
&= \lim_{\delta \to 0} \frac{F_X\big(g^{-1}(y+\delta)\big) - F_X\big(g^{-1}(y)\big)}{g^{-1}(y+\delta) - g^{-1}(y)} \, \frac{g^{-1}(y+\delta) - g^{-1}(y)}{\delta} \\
&= \lim_{\delta,\, \epsilon \to 0} \frac{F_X\big(g^{-1}(y) + \epsilon\big) - F_X\big(g^{-1}(y)\big)}{\epsilon} \, \frac{g^{-1}(y+\delta) - g^{-1}(y)}{\delta} \\
&= \frac{\mathrm{d}F_X(t)}{\mathrm{d}t} \bigg|_{t = g^{-1}(y)} \, \frac{\mathrm{d}g^{-1}(s)}{\mathrm{d}s} \bigg|_{s=y} \\
&= f_X\big(g^{-1}(y)\big) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}
\end{aligned}$$

Or more succinctly,

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{\mathrm{d}F_X\big(g^{-1}(y)\big)}{\mathrm{d}g^{-1}(y)} \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} = f_X\big(g^{-1}(y)\big) \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}$$

Imagine Y goes from y to $y + \delta$, accumulating probability $P(y < Y < y + \delta)$. This probability is equal to that accumulated by X if X goes, correspondingly, from $g^{-1}(y)$ to $g^{-1}(y + \delta)$.

If we take the ratio of the probability to the distance traveled by Y and X, respectively, we get the pdf of Y and X. These two densities differ, because the distances traveled by Y and X differ. How do we find out the (relative) difference in the traveled distance? That is related to the slope of the transformation—the derivative of g or g^{-1}.

Suppose $g(x) = 2x$. Then Y travels twice as far as X. Hence the density of Y is half that of X. Note $g^{-1}(y) = \frac{y}{2}$ and $\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} = \frac{1}{2}$.

Of course, when the derivative of g is not constant, we'll check the relative pace of X and Y in a local context, that is, near "corresponding" locations: Y at y and X at $g^{-1}(y)$.

How can I remember this formula?

We know the pdf of X and we need to get the pdf of Y. So we need to transform Y to X and then use the pdf of X. Note there is only <u>one function</u> involved: the function that transforms Y to X, i.e. g^{-1}. We calculate the pdf of X at $g^{-1}(y)$. Hence the first term. We take the slope of this transformation. Since it's a function of Y, then the derivative has to be with respect to Y. Hence the second term.

Example 2.1.6 (page 51).

Example 2.1.7 (page 52). Revisited.

Here the function g is not monotone. We're not using the theorem 2.1.5 (which is not usable). Instead we work out the cdf by direct reasoning. Then we get the pdf by simply differentiating the cdf.

However, this example shows the essence of the theorem 2.1.5: we know the pdf as long as we know the cdf. When g is monotone, the transformation between X and Y is one-to-one and on-to, hence there is a correspondence in probability. Therefore we can express the cdf of Y in terms of the cdf of X.

This idea generalizes in theorem 2.1.8: if we can partition $\mathcal{Y}$ such that in each part we can work out the probability $P(Y \leq y)$, then the cdf of Y is a summation. After that, the pdf of Y is straightforward to derive.

Theorem 2.1.8 (page 53): generalization of theorem 2.1.5.

Understand the conditions listed in this theorem.

Understand a further generalization if condition iii is not satisfied.

Example 2.1.9 (page 53).

Theorem 2.1.10 (page 54). This theorem provides a general method for sampling from a distribution, as long as we know the cdf (and the inverse of it).

In practice, however, simulating from a common distribution often makes use of relations between distributions. Simulate from an easier-to-simulate distribution, then a certain transformation the simulated values gives a sample from another distribution.

Uniform is the most thoroughly studied distribution as far as "random number generators" are concerned.

Example 2.1.4 (page 51) revisited. Can you propose a way to simulate an exponential variable?

Example 2.1.9 (page 53) revisited. Can you propose a way to simulate a chi-square variable?

Example 2.1.6 (page 51) revisited. Can you propose a way to simulate an inverted gamma variable?

5 Expected values

The definition of expected values should be already familiar to anyone who is taking this course.

Perhaps only one thing is different from the definition in an intro course: we define the expected value of a random variable $g(X)$ directly, not just X.

Conventionally, the definition describes the discrete and continuous cases separately.

Definition 2.2.1 (page 55).

Remarks 1. Do not misunderstand the "expected" value as the "most probable" value or a "typical" value. Unless the distribution is symmetric, $E(X)$ is not the point where the pdf or pmf peaks. Example: pdf of the exponential function peaks at 0 while its expected value is λ.

2. $E(X)$ may be a value that the random value can never take. For example, $E(X)$ of a discrete X is more often than not a value outside of the set of possible values of X.

3. $E(X)$ may not exist. This is the case with some distributions with heavy tails. Note the absolute value symbol: if $E|g(X)| = \infty$, we say $E(g(X))$ does not exist.

4. Commonly used symbol: μ. (A good alternative, if μ is not usable, is m.)

Example 2.2.4 (page 56): $E(X)$ of Cauchy distribution does not exist.

Properties Theorem 2.2.5 (page 57).

All properties are intuitive. The most important of the properties is the linearity of the E operator: $E(aX+b) = aE(X)+b$.

This can be built up from several more basic properties:

1. $E(c) = c$. (Expected value of a constant is itself.)
2. $E(cX) = cE(X)$.
3. $E(X+Y) = E(X) + E(Y)$. (This does NOT require independence b/t X and Y!)
4. $E(X+c) = E(X)+E(c) = E(X)+c$. (This is a special case of the above.)

Note $E(XY) \neq E(X)\,E(Y)$.
$E\big(g(X)\big) \neq g\big(E(X)\big)$.

6 Moments and variance

Definition 2.3.1 (page 59): moments and central moments.

Definition 2.2.2 (page 59): variance and standard deviation.

Note 1. Interpretation of variance: spread around its mean. (Due to this meaning, the variance of an asymmetric distribution such as the exponential, is somewhat less useful.

2. Units of variance and standard deviation.

3. Commonly used symbols: σ^2 and σ. (An alternative is V.)

4. The first two moments are extremely important. The 3rd and 4th are occasionally used. Forget about higher-order moments.

Properties 1. $\operatorname{var}(aX+b) = a^2 \operatorname{var}(X)$. (theorem 2.3.4, page 60)
2. $\operatorname{var}(X) = E(X^2) - \big(E(X)\big)^2$.

Prove both.

Note 1. $\operatorname{var}(c) = 0$. (Variance of a constant is 0.)

2. $\operatorname{var}(X+c) = \operatorname{var}(X)$. (Adding a constant does not change the variance.)

3. $\operatorname{var}(aX) = a^2 \operatorname{var}(X)$. (Multiplying a constant gets squared in variance.)

4. $\operatorname{var}(X+Y) \neq \operatorname{var}(X) + \operatorname{var}(Y)$ (unless X and Y are independent).

7 Examples

Example 2.2.2 (page 55).

Example 2.3.3 (page 59). You should master the technique of "integration by parts" used in this example. It's a recurrent pattern.

Example 2.2.4 (page 56).

Example 2.3.5 (page 61).

8 Recollection: what distributions have we encountered?

What do you know about each: cdf? pdf or pmf? mean? variance? support? relation in between? Make a list of the examples that addresses each of these distributions; study the examples.

uniform

Bernoulli (toss of a coin; two outcomes)

Binomial

geometric

normal

χ^2

exponential

Cauchy

9 Moment generating functions

"Moment generating function" is one of several "generating" functions in probability. Although a mgf can "generate" moments, its main use is not in calculating moments, but rather in characterizing (i.e. identifying) a distribution. However, it is not particularly easy to use for that purpose.

Definition 2.3.6 (page 62).

Note 1. mgf is defined if the expectation Ee^{tX} exists (and is necessarily finite) in some neighborhood of 0. We only need this function (of t) in the neighborhood of 0 (see example 2.3.8), because we'll take its derivatives at 0 (see theorem 2.3.7). For this purpose, the neighborhood can be arbitrarily small as long as it's a real interval that contains 0.

2. We've seen that the mean is an expected value, the moments are expected values, now the mgf is an expected "function".

3. A usual notation is $M_X(t)$.

4. The theorems 2.3.7 and 2.3.11 show how to use mgf.

Theorem 2.3.7 (page 62). Calculating moments via mgf. Prove it.

Example The gamma pdf is given by

$$f(x) = \frac{1}{\gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty;\ \alpha, \beta > 0.$$

(1) Find EX directly. (Answer: $\alpha\beta$.)
(2) Find the mgf. (Example 2.3.8, page 63.)
(3) Find EX via the mgf.

Theorem 2.3.11 (page 65).

Note 1. If two distributions have identical sequence of moments, the two distributions may not be the same. Part (a) says they are if in addition they both have bounded support. (However, it is very hard to verify that two distributions have identical sequence of moments. Therefore this is not a useful way to identify distributions.)

2. Part (b) is the much more important part. It says mgf uniquely identifies a distribution (if the mgf exists).

3. We've learned that cdf uniquely identifies a distribution. Now mgf is another such thing. (However, whereas cdf always exists, mgf may not.) We can almost take cdf as the definition of a distribution.

4. Theorem 2.3.12 is less important for us, because (1) We probably won't need to use mgf that skillfully; (2) We have yet to learn the concept of "convergence" of random variables or distributions.

Theorem 2.3.15 (page 67). A property of mgf. Prove it.

Lemma 2.3.14 (page 67). A very useful calculus result. Another useful one is the sum of a geometric series, (1.5.4) on page 31.

Example The Poisson pmf is given by

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots;\ \lambda > 0.$$

Find the mgf. (Answer: $e^{\lambda(e^t-1)}$.)

Example The binomial pmf is given by

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n;\ 0 < p < 1.$$

(1) Find EX directly. (Example 2.2.3, p. 56.)
(2) Find the mgf. (Answer: $\big(pe^t + (1-p)\big)^n$. Example 2.3.9, p. 64.)
(3) Find EX and EX^2 via the mgf.

To take the derivatives of $M_X(t)$, notice $\log M_X(t) = n \log\big(pe^t + (1-p)\big)$. Then

$$\frac{\mathrm{d}M_X(t)}{\mathrm{d}t} = M_X(t) \frac{\mathrm{d}\log M_X(t)}{\mathrm{d}t} = M_X(t)\, npe^t (pe^t + 1 - p)^{-1}$$

To find the second derivative, continue from this first derivative and use the chain rule. This is a useful trick.

Example The exponential pdf is given by

$$f(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x > 0;\ \lambda > 0.$$

(1) Find the mgf. (Answer: $\frac{1}{1-\lambda t}$.)
(2) Find EX, EX^2, and $\operatorname{var}(X)$. (Answer: λ, $2\lambda^2$, λ^2.)

Comment Exponential is a special case of gamma.

Example The pdf of standard normal is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Find the mgf. (Answer: $e^{t^2/2}$.)

Example The pdf of normal $N(\mu, \sigma^2)$ is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}.$$

(1) Find the mgf. (Hint: theorem 2.3.15. Answer: $e^{\mu t + \frac{1}{2}\sigma^2 t^2}$.)
(2) Find EX, EX^2, and $\operatorname{var}(X)$. (Hint: use the $\log M_X(t) = \cdots$ trick.)

Example We have seen in Example 2.1.9 (p. 53) that the pdf of χ^2 with one degree of freedom is

$$f(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2}.$$

Find the mgf.

Comment Chi-square is a special case of gamma.

Example Suppose a random variable X takes the possible values 0, 1, and 2 with respective probabilities 1/8, 1/4, and 5/8. The mgf is then obviously $M_X(t) = \frac{1}{8} + \frac{1}{4}e^t + \frac{5}{8}e^{2t}$.

Example We have a random variable whose mgf is $M_X(t) = \frac{1}{5}\big(1 + e^t + 3e^{2t}\big)e^{-t}$. What's the pmf of X? (Hint: what are its possible values, and what are the respective probabilities?)

Example Suppose $M_X(t) = \frac{1}{16}(1 + e^t)^4$. What's the distribution of X? (Hint: compare with the mgf of binomial.)

Example Suppose $M_X(t) = e^{\pi t^2}$. What's the distribution of X? (Hint: compare with the mgf of normal.)